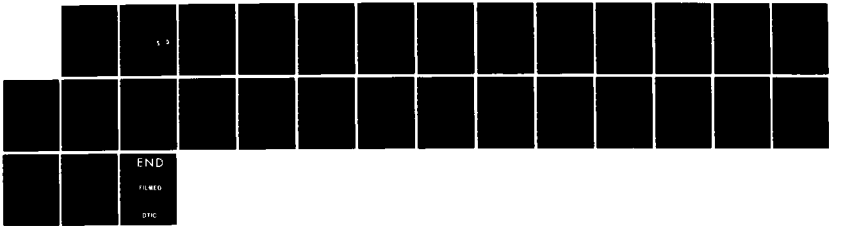
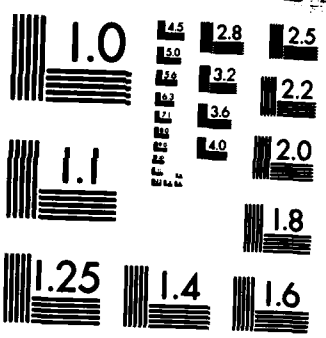
AD-A153 396 ON SEGMENTATION OF SIGNALS TIME SERIES AND IMAGES(U) 1/1
ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE
METHODS S L SCLOVE 01 MAR 85 UIC/DQM/ARO-84-2
UNCLASSIFIED ARO-19085.11-MA DAAG29-82-K-0155 F/G 12/1 NL





UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(2)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 19085.11-MA	2. GOVT ACCESSION NO. N/A	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) On Segmentation of Signals, Time Series, and Images		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stanley L. Sclove		8. CONTRACT OR GRANT NUMBER(s) DAAG29-82-K-0155
9. PERFORMING ORGANIZATION NAME AND ADDRESS Quantitative Methods Department University of Illinois at Chicago Box 4348, Chicago, IL 60680-4348		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE March, 1985 13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

statistical pattern recognition, classification; temporal correlation, spatial correlation; optimization by relaxation method

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Signals and time series often are not homogeneous but rather are generated by mechanisms or processes with various phases. Similarly, images are not homogeneous but contain various objects. "Segmentation" is a process of attempting to recover automatically the phases or objects. A model for representing such signals, time series, and images is discussed. Some approaches to estimation and segmentation in this model are presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARO 19085.11-MP

ON SEGMENTATION OF SIGNALS, TIME SERIES, AND IMAGES

Stanley L. Sclove

A presentation to the
30th Conference on Design of Experiments
in Army Research, Development, and Testing

October 17-19, 1984

TECHNICAL REPORT NO. UIC/DQM/ARO 84-2
March 1, 1985

PREPARED FOR THE
ARMY RESEARCH OFFICE
UNDER
CONTRACT DAAG29-82-K-0155

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348, CHICAGO, IL 60680

3/9/85

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

ON SEGMENTATION OF SIGNALS, TIME SERIES, AND IMAGES

Stanley L. Sclove

Department of Quantitative Methods
University of Illinois at Chicago

CONTENTS

ON SEGMENTATION OF SIGNALS, TIME SERIES, AND IMAGES

Stanley L. Sclove

Department of Quantitative Methods
University of Illinois at Chicago

ABSTRACT

Signals and time series often are not homogeneous but rather are generated by mechanisms or processes with various phases. Similarly, images are not homogeneous but contain various objects. "Segmentation" is a process of attempting to recover automatically the phases or objects. A model for representing such signals, time series, and images is discussed. Some approaches to estimation and segmentation in this model are presented.

Key words and phrases: statistical pattern recognition, classification; temporal correlation, spatial correlation; optimization by relaxation method.

ON SEGMENTATION OF SIGNALS, TIME SERIES, AND IMAGES

Stanley L. Sclove

Department of Quantitative Methods
College of Business Administration
University of Illinois at Chicago

1. Introduction

About ten years ago, Professor James Osterburg, a colleague at the University of Illinois at Chicago and an expert on physical evidence (Osterburg and O'Hara, 1949; Osterburg, 1968, 1982), consulted me in regard to establishing probability estimates for partial fingerprints. We have had a very interesting collaboration which resulted in several papers (Osterburg, Parthasarathy, Raghavan, and Sclove, 1979; Sclove, 1979, 1980, 1981a) and may, thanks to Professor Osterburg's continuing efforts, result in changes in practices regarding the evaluation of fingerprints as evidence. For the present paper the point about fingerprints is this: When at professional meetings I would talk about the subject, certain people (usually electrical engineers or computer scientists) would tell me that what I was doing was "image processing."

The reason that our fingerprint work resembles image processing is this. Osterburg treated fingerprint analysis by placing a grid of cells over the print. One then notes the locations of any occurrences of the "Galton details," the minutiae of the ridge lines, such as ridge endings and forks. In numerical image processing a real image is divided into cells ("pixels": picture elements) and one notes numerically what occurs in each cell. The real image is expressed as a matrix -- rows by columns -- of cells, just as the TV screen has a matrix of dots which are illuminated with various colors and intensities.

2. Images

An important aspect of image processing is image "segmentation," the location of objects in the image.

Examples. (i) A picture of a house and yard is to be labeled using the labels, brick, glass, tree, grass, sky. (ii) An image is to be labeled using the labels, tank, mud, tree, sky. (iii) A medical image is to be labeled using the labels, tumor, or normal tissue.

Segmentation involves labeling each pixel according to the name of the object of which it is a part, in turn involves the grouping of neighboring pixels. Since I had earlier worked on cluster analysis, the grouping of observations (Sclove 1977), it was natural to concentrate on segmentation. The labeling process can be made explicit in a model which states that at each pixel we observe the value of a random variable X, but also along with X there is an unobservable variable, the label. (I tend to treat the labels as parameters. Others treat them as missing data, i.e., random variables.) In the context of this model, segmentation is merely estimation of the labeling parameters.

The random variable X is often a vector of several measurements.

Examples. (i) A familiar example of a vector of measurements (though it would not be measured across a two-dimensional array) is blood pressure, which is a vector of the two measurements, systolic and diastolic. (ii) In color television, X is a vector of three measurements, the red level, green level, and blue level. (iii) In Landsat satellite data, there are four spectral channels, one in the green/yellow visible range, a second in the red visible range, and the other two in the near infrared range. (iv) For a black-and-white image, the random variable X which is simply a scalar consisting of a single gray-level measurement, rather than a vector.

Images are two-dimensional; I decided first to consider a one-dimensional version of the problem. An image is a two-way series; a one-way series is a "time series." So I began this research by thinking about segmentation of time series.

3. Time Series

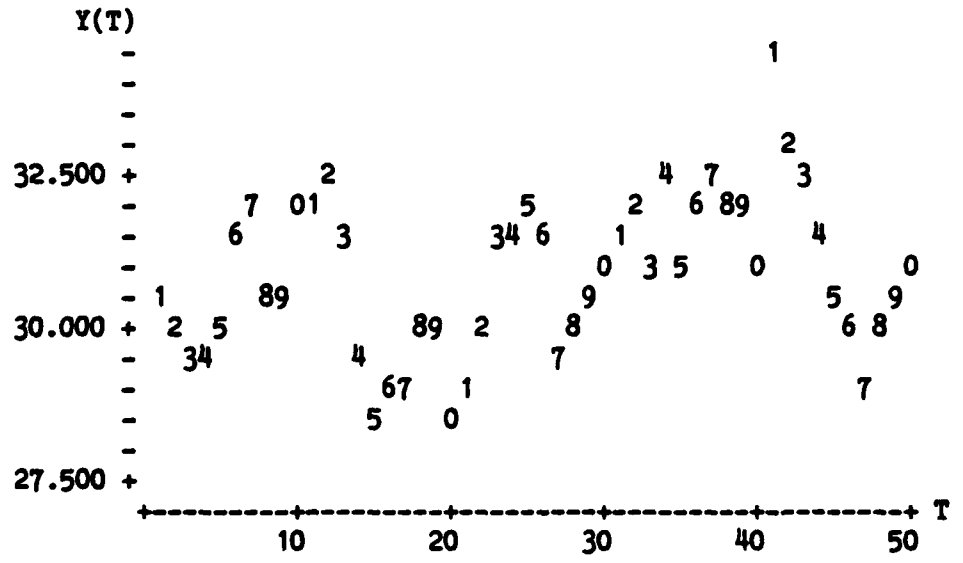
The problem of segmentation considered here is: Given a time series

$$\{x(t),\ t=1,2,\ \ldots\ ,n\},$$

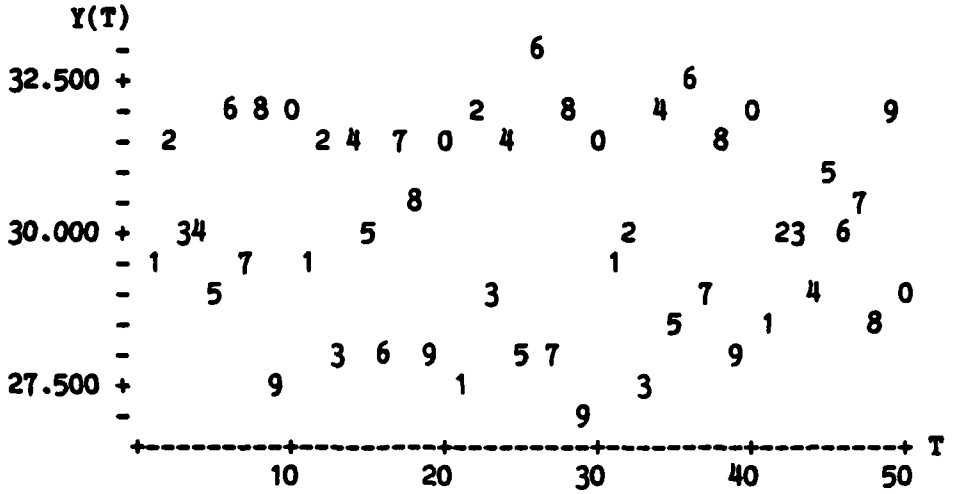
partition the set of values of t into subseries (segments, regimes) for which the values x(t) are relatively homogeneous in some sense. The segments are assumed to fall into several classes.

Examples. (i) Segment a received signal into background, target, background again, another target, etc. (ii) Segment an EEG of a sleeping persion into periods of deep sleep and restless or fitful sleep (two classes of segment). (iii) Segment an ECG into rhythmic and arhythmic periods (two classes of segment). (iv) Segment an economic time series into periods of recession, recovery, and expansion (three classes of segment).

Next, several simulated time series will be shown. The first is relatively smooth; the second, relatively rough or jumpy. (They are simulated first-order autoregressions with autoregression coefficients equal to +.8 and -.8, respectively.)

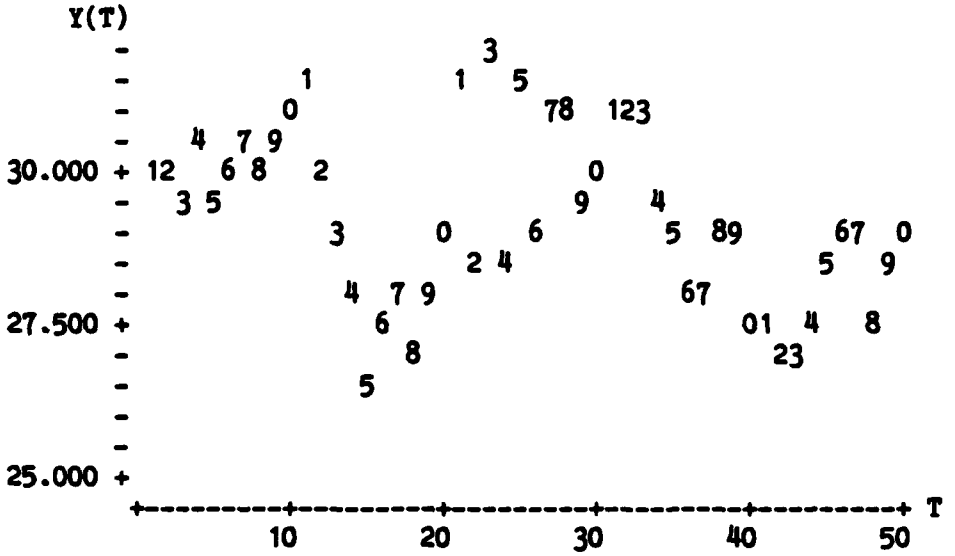


Smooth time series (simulated first-order autoregression with coefficient +.8)



Rough time series (simulated first-order autoregression with coefficient -.8)

The next series has alternating smooth and jumpy segments.

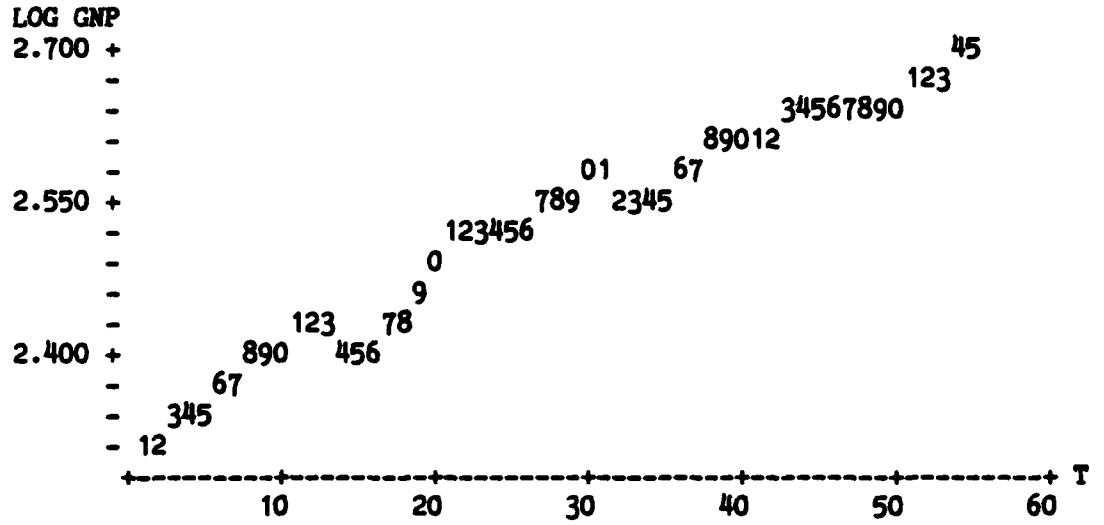


Simulated time series with three segments (smooth, then rough, then smooth)

It is a simulated series, with three segments. Each segment is a first-order autoregression. For t = 1 to 20 the coefficient is +.8; for

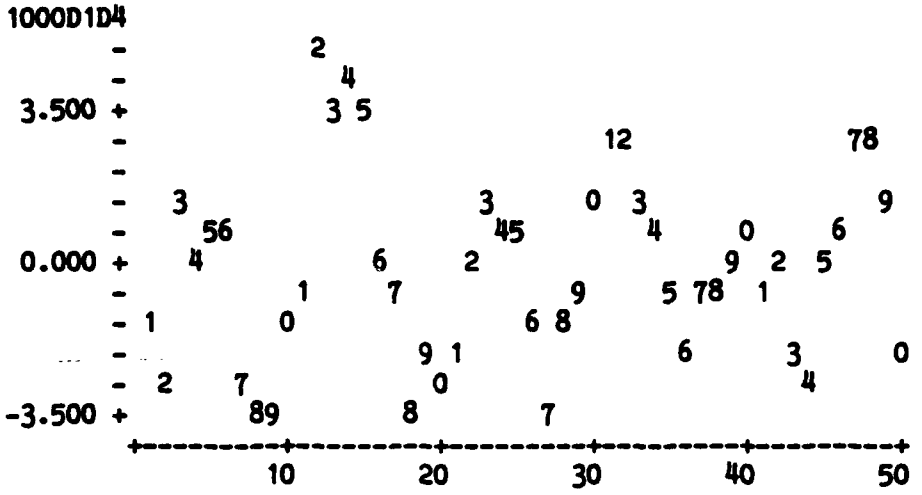
t = 21 to 30 the coefficient is -.8; then for t = 31 to 50 the coefficient is again +.8. In actual data analysis one would not know the segments but would have to find them and estimate their parameters.

Now let us consider a real example, rather than simulated data. The graph below is quarterly GNP (Gross National Product) for 55 quarters, starting with the first quarter of 1946.



Plot of Y(T), where Y(T) is log (base ten) quarterly GNP in billions of current (unadjusted) dollars. T=1 is 1946-1, T=2 is 1946-2, ..., T=5 is 1947-1, etc. (N=55)

Here is a second difference of the log quarterly GNP. Do you see anything particularly interesting?



Plot of mixed second difference of log GNP: Vertical scale, labeled '1000D1D4', is Z(T), where Z(T) = 1000[(Y(T)-Y(T-1)) - (Y(T-4)-Y(T-5))] and Y(T) = log GNP (base 10); T=1 is 1947-2, T=2 is 1947-3, ..., T=9 is 1949-2, etc. (N=50)

A first difference corresponds to a velocity. The difference

$$y(t) - y(t-1) = d(t),$$

say, is $[y(t) - y(t-1)]/[t-(t-1)]$, which is the change in y over the one time unit from time t-1 to time t. A second difference

$$d(t) - d(t-1),$$

which in terms of an original series of y's is

$$\{[y(t)-y(t-1)] - [y(t-1)-y(t-2)]\},$$

is proportional to the change in velocity across the indicated time periods and hence is essentially an acceleration. Thus a second difference is perhaps a natural transform to analyze. (Note, however, that the second difference used here is a mixed second difference, namely, the ordinary difference of the lag-four difference y(t)-y(t-4).] Ths second-difference series appears level. What I think is particularly interesting is that the acceleration of the economy due

to the Korean conflict of the 1950's is readily apparent. (T = 12 is 1950-1, T = 13 is 1950-2, etc.) I think that the need for segmentation is clearly indicated. One needs to give some sort of special treatment to those four exceedingly high values.

An alternative approach to such observations is to identify them as "outliers." However, "outlier" connotes spuriousness. If outlying observations are non-spurious or are associated with a recurring cause, perhaps they should not then be termed "outliers." They should be modeled.

In other cases, where, e.g., equipment failure is suspected, one truly wants to look for outliers. Here, too, segmentation can be useful.

De Alba and Zartman (1979) analyzed radiotelemetric measurements of cows' temperatures, in order to locate the periods of high estrus, with a view toward more optimally timed breeding and efficient milk production. A technique of de Alba and Van Ryzin (1979, 1980) was used. In the de Alba-Zartman report the analysis of the temperatures of one cow over 133 days is discussed in detail. Eleven observations were detected as coming from distributions with means shifted relative to the rest of the observations. Inspection of the data showed that some of these temperatures were high and some extremely low. It is concluded that the high temperature readings correspond to times of active estrus. (Perhaps the low readings correspond to instrument failure and hence are true "outliers" in that sense.)

Here, as in de Alba and Zartman (1979), the raw data were pre-processed by differencing. Model-selection criteria (see below)

estimated the number, k, of classes of segment as two, but k = 3 and k = 6 scored almost as well. Since the numbers analyzed were differences, it is perhaps especially interesting to consider three classes, particularly since the segmentation yielded one class for positive differences, one for negative differences, and one for differences close to zero. The results obtained are similar to those obtained by de Alba and Zartman. They identified eleven observations as "outliers;" 7 of these were high and 4 were low. The segmentation algorithm run with three classes of segment identified 4 observations as high; these were among the 7 identified by de Alba and Zartman. Use of six classes provided even closer agreement to the de Alba-Zartman results. The upper two of the six classes captured 6 of the 7 high observations; the 4 low observations were located by the lowest class.

(In discussion after the paper Professor Parzen mentioned that at Texas A&M University also they had dealt with the analysis of bovine estrus and found by spectral analysis that a filtered Poisson process -- see, e.g., Parzen (1962) -- provided a good fit to the data.)

Now, having discussed some examples, let me be a little more specific about the model and the algorithm. [More formal presentations are found in Sclove (1983a,c,d).] The elements of the segmentation model are the class-conditional time series or distributions, with their parameters; the labels; and the transition probabilities between the labels. Correspondingly, the algorithm alternates between estimation of the distributional parameters, estimation of the labels, and estimation of the transition probabilities. That is, given a tentative labeling,

one can obtain tentative estimates of the parameters of the class-conditional distributions and of the transition probabilities. One then relabels the observations, using these updated parameter estimates. The relabeling is done as follows: If x(t) (i.e., if time period t) is currently labeled as class c, then x(t+1) is labeled as that class d for which the product

$$p(c \text{ to } d)f(x(t+1)|d)$$

is maximal, where p(c to d) denotes the current estimate of the probability of a transition from c to d and f(x|d) denotes the tentatively estimated class-d probability density, evaluated at x. This makes sense because, under the assumptions of the model, the likelihood is the product of these terms.

To illustrate the algorithm, let us consider a short, artificial time series.

t:	1	2	3	4	5	6	7	8	9	10	11	12
x(t):	1	1	3	1	2	1	2	6	7	1	1	1

Suppose it is specified that there are two classes and that the class-conditional distributions are exponential. Suppose the initial guesses of the parameters are equal prior probabilities for the two classes and means of 2 and 3. Then initially the class-conditional densities are taken as $f(x|a) = (1/2)\exp(-x/2)$ and $f(x|b) = (1/3)\exp(-x/3)$. In the first iteration, using equal prior probabilities of .5, one labels x as having come from Class A if $.5f(x|a) > .5f(x|b)$, which simplifies to $x < 2.43$. This gives the following estimated labels at the end of the first iteration.

t:	1	2	3	4	5	6	7	8	9	10	11	12
x(t):	1	1	3	1	2	1	2	6	7	1	1	1
label:	a	a	b	a	a	a	a	b	b	a	a	a

Now the transition probabilities can be estimated from the sequence of estimated labels. The 11 transitions are: a to a, a to b, b to a, a to a, a to a, a to a, a to b, b to b, b to a, a to a, a to a. These give the following estimates of the transition probabilities.

$$p(a \text{ to } a) = 3/4 \qquad p(a \text{ to } b) = 1/4$$

$$p(b \text{ to } a) = 2/3 \qquad p(b \text{ to } b) = 1/3$$

The class means are estimated as follows (but see Section 6: Plans for Further Research, for a discussion of this point).

Mean of a's: $(1 + 1 + 1 + 2 + 1 + 2 + 1 + 1 + 1)/9 = 11/9 = 1.22$

Mean of b's: $(3 + 6 + 7)/3 = 16/3 = 5.33$

Now the condition for labeling the current x as "a", given that the preceding x has been labeled as "a", is

$$p(a \text{ to } b)f(x|b) < p(a \text{ to } a)f(x|a),$$

or

$$(1/4)(1/5.33)\exp(-x/5.33) < (3/4)(1/1.22)\exp(-x/1.22),$$

which simplifies to $x < 4.075$. Similarly, the condition for labeling x as "a", given that the preceding observation has been labeled as "b", is $p(b \text{ to } b)f(x|b) < p(b \text{ to } a)f(x|a)$, which simplifies to $x < 1.24$. These second-iteration classification rules change only the label of x(3) from "b" to "a".

At this point let me mention a possible improvement to the algorithm. At the relabeling stage, where the labels are re-estimated, based on tentative parameter estimates and transition probabilities, the

problem is one of estimating a finite sequence, given the transition probabilities. The dynamic programming approach to this problem is known as the Viterbi algorithm. [Forney (1971) is perhaps the most readable reference on this; the references Viterbi (1967, 1971) and Viterbi and Odenwalder (1969) are also given.] This approach will be applied in the future. The present approach is perhaps somewhat simpler but has its advantages. It adapts itself to operation in the purely sequential mode, and it is relatively easy to program.

An algorithm, such as the present one, which alternates between optimizing different sets of variables, is known as a "relaxation" method (Southwell 1940, 1946; Ortega and Rheinboldt 1970). In the present case the different sets of variables are different parameters, namely, the distributional parameters, the transition probabilities, and the labels.

We note in passing that the EM algorithm -- see Dempster, Laird and Rubin (1977) -- is such a relaxation method, where at the "E" step the estimation is by expectation and at the "M" step the estimation is by maximization.

An alternative approach to the presently-implemented relaxation method would involve equating partial derivatives of the likelihood function to zero and solving the resulting equations. It is clear that results analogous to those of Wolfe (1970) will be obtained by this approach, with transition probabilities replacing his mixture probabilities. In any case, the resulting equations have to be solved by an iterative, numerical method, and it is not clear whether it would be any better than the the presently-implemented relaxation method.

4. Image-segmentation experiments

Conventional approaches to segmentation (see, e.g., Ahuja and Rosenfeld (1982) for a survey of image models) include ordinary clustering; edge detectors such as the two-dimensional filters of Irwin Sobel or Judy Prewitt; and a pixel-labeling method that has (perhaps inaccurately) been called a relaxation method. This involves updating the current estimate of

$$\Pr(\text{label of } t \text{ is } c \mid \text{data}),$$

the (conditional) probability that the true label of pixel t is class c, given the data, by an updated version of these estimated probabilities, where the updating takes into account the current labels of neighboring pixels. This is done by means of "compatibility coefficients," measuring the consistency of label c for pixel t with the current labelings of neighboring pixels. (See, e.g., Eklundh, Yamamoto and Rosenfeld 1980.) More precisely, to estimate $\Pr(\text{label of } t \text{ is } c \mid \text{data})$, one moves from stage s-1 to stage s as follows. Let

$$\Pr(\text{label of } t \text{ is } c \mid \text{data}; s)$$

be the s-th stage estimate. Then

$$\Pr(\text{label of } t \text{ is } c \mid \text{data}; s) = \Pr(\text{label of } t \text{ is } c) \mid \text{data}; s-1)\, C(t,c;s-1)/C,$$

where C is a normalizing factor equal to the sum of the denominators in this expression and C(t,c;s-1) is the compatibility coefficient; e.g., C(t,c;s-1) could be related to estimated transition probabilities for the labels or taken as the proportion of the neighbors of pixel t which are labeled c at stage s-1. A problem with this relaxation

method, noted by various researchers, is that the image gets good, then "fades," so there is a problem of knowing when to stop the iteration.

This approach is intuitively attractive. The approach I use seems to have the same intuitive attraction, plus the advantage of being embedded in a Markov model which gives a likelihood function, putting the problem in the context of parameter estimation in a probability model and making performance evaluation possible. (For instance, one can consider asking questions such as, "What if we worked with first-order neighbors and the model were really second-order?" I'm not saying that such questions are easy to answer; I'm merely stating that it is only in the presence of a model that their formulation is even possible.)

The idea of the model in the case of images is the same as for time series, but the transition matrix is more complicated. The transition probabilities are, even in the simplest, first-order, one-sided case (where one conditions only on pixels to the north and west of the given pixel, rather than those to the north, west, south, and east), functions p((c,d) to e) of three arguments, where this represents the probability of a transition to class e in pixel t, given that the pixel to the north of t is class c and the pixel to the west of t is class d. [The Markov approach used here is not unrelated to that presented by Professor Grenander at this conference (Grenander 1985; see also Grenander 1983).] The segmentation algorithm discussed in the present paper, like the "relaxation" method using compatibility coefficients, also has the property that each iteration is not necessarily better than the preceding, but there is a likelihood value associated with each

iteration, and these values can be used to pick out the best iteration.

Here are results of some experiments with the segmentation algorithm.

The Fisher iris data consist of 4 variables observed for each of 150 irises, 50 in each of three species. In order to form an experimental image, these 150 were arranged into 15 rows of 10, the first 5 rows being species 1, rows 6-10 being species 2, and rows 11-15 being species 3. Thus the true segmentation looks like this.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

The measurements on flowers 50, 100, and 150 were used as initial estimates for the distributional parameters, namely, means of multivariate normal distributions. A common covariance matrix was assumed. (Actually, statistical tests and model-selection criteria suggest that different covariance matrices should be used; my algorithms do allow for this.) The initial estimate of the common covariance was taken to be proportional to the identity matrix. Equal prior probabilities for the three classes were assumed as initial estimates. After only three iterations, no label changed. The segmentation

obtained was as follows.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
2 3 2 3 2 3 2 3 2 3
2 2 2 2 2 2 2 2 2 2
3 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

Experiments were run with different configurations of these data. Below is the true configuration for Experiment 1. The four-by-four block of 2's in the middle can be thought of as a target to be located.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 2 2 2 2 1 1 1
1 1 1 2 2 2 2 1 1 1
3 3 3 2 2 2 2 3 3 3
3 3 3 2 2 2 2 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

The segmentation produced was as follows.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 3 2 3 2 1 1 1
1 1 1 3 2 2 2 1 1 1
3 3 3 2 2 2 2 3 3 3
2 3 3 2 2 2 2 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

The following was the true configuration for Experiment 2. Note the small "target" of four 2's.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 2 2 1 1 1 1
3 3 3 3 2 2 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

The segmentation obtained was as follows.

```
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 3 2 1 1 1 1
3 3 3 3 3 2 3 3 2 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3
```

Note the algorithm worked reasonably well in detecting even a small target, even with the same initial (equal) estimates of prior probabilities, quite different from the true probabilities of .48, .04, .48 for the three classes. Thus perhaps the algorithm will not have to be well "tuned" in order to work well.

Note there are problems at the northwest edges, due to the way the algorithm is set up. In future software development I might program a procedure analogous to Box and Jenkins' back-forecasting ("backcasting") to take care of this. Then, after a first segmentation, the data would be read through in reverse order, so that northwest pixels become southeast pixels, easy to label correctly.

5. Model-selection criteria

A question of obvious importance is that of how many classes of segment to use, that is, what should be the value of the parameter k? Maximum likelihood estimation alone cannot provide an answer, because it is model-conditional; that is, maximum likelihood applies to the problem of estimating the other parameters, for a fixed value of k. It does not apply to the problem of estimating k because the likelihood itself changes as k does. An approach to the problem of estimation of k is provided by model-selection criteria, such as Akaike's, Schwarz' and Kashyap's. These criteria add a penalty for using extra parameters to the negative log-likelihood function. Thus a small score is a good score on these criteria. Akaike's criterion is based on a heuristic estimate of the cross-entropy of the true model and the model with k classes. (See Parzen, 1982.) Schwarz' criterion, perhaps more convincing, is based on a Bayesian approach. Kashyap obtains Schwarz' criterion by a relatively simple expansion, then takes this expansion a term further to obtain another criterion, which I call Kashyap's criterion. The additional term seems to be particularly meaningful. These model-selection criteria take the form

$$-2 \log L(k) + a(n)m(k) + b(k),$$

where "log" here denotes the natural (base-e) logarithm, and

$L(k)$ = likelihood under the k-th model, maximized with respect to the parameters,

$a(n) = 2$, for all n, $b(k) = 0$, for Akaike's criterion,

$a(n) = \log n$, $b(k) = 0$, for Schwarz' criterion,

$a(n) = \log n$, $b(k) = \log[\det B(k)$, for Kashyap's criterion,

det = determinant,

and

B(k) = matrix of second partial derivatives of log L(k) with respect to the parameters.

(The mathematical expectation of B(k) is the Fisher information matrix.) The optimal k is that value which minimizes the criterion. Akaike's criterion generally chooses a higher value of k (more parameters) than do the others. Since for n greater than 8, log n is greater than 2, Schwarz' criterion will choose a value of k no larger than that chosen by Akaike's, for n greater than 8.

6. Plans for further research

Several items for future research have already been mentioned, including programming of backcasting and use of the Viterbi algorithm.

Some other plans for additional research will be mentioned now. First there is the matter of improved estimation of the distributional parameters. For purposes of discussion focus on the example of segmenting a time series into two classes, i.e., labeling each observation as an "a" or a "b." There is a truncation resulting from the present implementation of the algorithm. Namely, only those observations labeled "a" will be used in updating the current estimate of the mean of Class A. But these observations are a truncated sample from Distribution A, and the algorithm does not treat them as such. (At the Conference, Professor Tukey very kindly sought me out the day after my presentation to point this out to me, and also to remark on the limitations of the unidirectional approach of the algorithm; see below.) Rather than deal with the truncation per se, I had planned in the next stage of the work to modify the estimators by doing them Bayesianly,

e.g., estimate the mean of Group A as N/D, where

$$N = x(1)\,Pr(a|x(1)) + x(2)\,Pr(a|x(2)) + \ldots + x(n)\,Pr(a|x(n))$$

and

$$D = [Pr(a|x(1)) + Pr(a|x(2)) + \ldots + Pr(a|x(n))]$$

(In this expression, $Pr(a|x)$ can be replaced by $Pr(x|a)$ since $Pr(a|x) = f(x|a)Pr(a)/f(x)$ and $Pr(a)/f(x)$ is a common factor which will cancel out.) In this estimate, all the observations play a role, whether labeled as "a" or "b," so that at least some of the bias will be removed by allowing the larger "b" observations to enter.

I obtain the likelihood function by a one-sided approach, conditioning any given pixel on the results in the pixels to its north and west. A two-sided, full neighborhood approach seems preferable to a unidirectional one. The unidirectional approach is a device for writing down the likelihood, but this does not mean one has to be wedded to that approach in the iterative updating. That is, the parameters can be estimated with a full neighborhood approach.

Another bit of further research is to calculate Kashyap's criterion for various clustering and segmentation models. Also, so far the method, algorithm and software have been developed only for the case where the observations within a class are independent (and Gaussian). A next step will be autoregression within classes. This is of obvious importance in time series, and in the context of images, it can be used to model textures. Still another generalization is to allow some forms of time or state dependency in the transition probabilities.

REFERENCES

Ahuja, N., and A. Rosenfeld (1982). Image models. Pp. 383-397 in Handbook of Statistics 2: Classification, Pattern Recognition, and Reduction of Dimensionality (P. R. Krishnaiah and L. N. Kanal, eds.), North Holland/American Elsevier, New York.

Akaike, H. (1983). Statistical inference and measurement of entropy. Pp. 165-189 in Scientific Inference, Data Analysis, and Robustness (H. Akaike and C.-F. Wu, eds.), Academic Press, New York.

Box, G.E.P., and G.M. Jenkins (1976). Time Series Analysis: Forecasting and Control, rev. ed. Holden-Day, San Francisco.

Bozdogan, H., and S. L. Sclove (1984). Multi-sample cluster analysis using Akaike's information criterion. Annals of the Institute of Statistical Mathematics 36, 163-180.

de Alba, E., and J. Van Ryzin (1979). An empirical Bayes approach to outliers. Journal of Statistical Planning and Inference 4, 217-236.

de Alba, E., and J. Van Ryzin (1980). An empirical Bayes approach to outliers: shifted mean case. Pp. 33-44 in Recent Developments in Statistical Inference and Data Analysis (K. Matusita, ed.), North Holland/American Elsevier, Amsterdam and New York.

de Alba, E., and Zartman, D.L. (1979). Testing outliers in time series: an application to remotely sensed temperatures in cattle. Special Paper Number 130, Agricultural Experiment Station, New Mexico State University, Las Cruces, New Mexico.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (B) 39, 1-38.

Eklundh, J. O., H. Yamamoto and A. Rosenfeld (1980). A relaxation method for multispectral pixel classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 2, 72-74.

Forney, G. D., Jr. (1973). The Viterbi algorithm. Proceedings of the IEEE 61, 268-278.

Grenander, U. (1983). Tutorial in pattern theory. Notes. Division of Applied Mathematics, Brown University, Providence, Rhode Island.

Grenander, U. (1985). Proceedings of this conference.

Johari, S., and S. L. Sclove (1976). Partitioning a distribution. Communications in Statistics A5, 133-147.

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Transactions on Pattern Analysis and Machine Intelligence 4, 99-104.

Ortega, J., and W. Rheinboldt (1970). Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York.

Osterburg, J. W., and O'Hara, C. E. (1949). An Introduction to Criminalistics. Macmillan, New York.

Osterburg, J. W. (1968). The Crime Laboratory: Case Studies of Scientific Criminal Investigation. Indiana University Press, Bloomington, Indiana.

Osterburg, J. W. (1982). The Crime Laboratory: Case Studies of Scientific Criminal Investigation, 2nd edition. Clark Boardman Co., Ltd., New York.

Osterburg, J. W., T. Parthasarathy, T.E.S. Raghavan, and S. L. Sclove (1977). Development of a mathematical formula for the calculation of fingerprint probabilities based on individual characteristics. Journal of the American Statistical Association 72, 772-778.

Parzen, E. (1962). Stochastic Processes. Holden-Day, San Francisco.

Parzen, E. (1982). Maximum entropy interpretation of autoregressive spectral densities. Statistics & Probability Letters 1, 7-11.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.

Sclove, S. L. (1977). Population mixture models and clustering algorithms. Communications in Statistics A6, 417-434.

Sclove, S. L. (1979). The occurrence of fingerprint characteristics as a two-dimensional process. Journal of the American Statistical Association 74, 588-595.

Sclove, S. L. (1980). The occurrence of fingerprint characteristics as a two-dimensional Poisson process. Communications in Statistics A9, 675-695.

Sclove, S. L. (1981a). Modeling the distribution of fingerprint characteristics. Pp. 111-130 in Statistical Distributions in Scientific Work 6 (C. Taillie et al., eds.). NATO Advanced Study Institutes Series. D. Reidel, Dordrecht, Holland

Sclove, S. L. (1981b). Pattern recognition in image processing using interpixel correlation. IEEE Transactions on Pattern Analysis and Machine Intelligence 3, 206-208.

Sclove, S. L. (1983a). Time-series segmentation: a model and a method. Information Sciences 29, 7-25.

Sclove, S. L. (1983b). Application of the conditional population-mixture model to image segmentation. IEEE Transations on Pattern Analysis and Machine Intelligence 5, 428-433.

Sclove, S. L. (1983c). On segmentation of time series. Pp. 311-330 in Studies in Econometrics, Time Series, and Multivariate Statistics (S. Karlin, T. Amemiya, and L. Goodman, eds.), Academic Press, New York.

Sclove, S. L. (1983d). On segmentation of time series and images in the signal detection and remote sensing contexts. Pp. 421-434 in Statistical Signal Processing (E. J. Wegman and J. G. Smith, eds.), Marcel Dekker, Inc., New York.

Southwell, R. (1940). Relaxation Methods in Engineering Science: a Treatise on Approximate Computation. Oxford University Press, London.

Southwell, R. (1946). Relaxation Methods in Theoretical Physics. Oxford University Press (Clarendon), London and New York.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory IT-13, 260-269.

Viterbi, A. J., and J. P. Odenwalder (1969). Further results on optimal decoding of convolutional codes, IEEE Transactions on Information Theory (Corresp.) IT-15, 732-734.

Viterbi, A. J. (1971). Convolutional codes and their performance in communication systems. IEEE Transactions on Communications Technology COM-19, 751-772.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329-350.

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348
CHICAGO, IL 60680

****

TECHNICAL REPORTS

ARMY RESEARCH OFFICE CONTRACT DAAG29-82-K-0155

with the University of Illinois at Chicago

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove

No. A82-1. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/82

No. A82-2. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-sample Cluster Analysis using Akaike's Information Criterion." 12/20/82

No. A82-3. Stanley L. Sclove. "Time-Series Segmentation: a Model and a Method." 12/22/82

No. A83-1. Hamparsum Bozdogan. "Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model using Model-Selection Criteria." 6/16/83

No. A83-2. Stanley L. Sclove. "On Segmentation of Digital Images using Spatial and Contextual Information via a Two-Dimensional Markov Model." Working Paper: 4/11/83; Technical Report: 12/6/83

No. A83-3. Stanley L. Sclove. "Use of Model-Selection Criteria in Clustering and Segmentation of Time Series and Digital Images." 5/5/83

No. A84-1. Stanley L. Sclove. "Pattern Recognition." 2/1/84

No. A84-2. Stanley L. Sclove. "On Segmentation of Signals, Time Series, and Images." 3/1/85. (Presented to the 30th Conference on Design of Experiments in Army Research, Development, and Testing, 10/17-19/84.)

No. A84-3. Hamparsum Bozdogan. "Multi-Sample Cluster Analysis as an Alternative to Multiple Comparison Procedures." 7/20/84

2/14/85

END

FILMED

5-85

DTIC